第27卷 第6期　　工 程 数 学 学 报　　Vol. 27 No. 6
2010年12月　　CHINESE JOURNAL OF ENGINEERING MATHEMATICS　　Dec. 2010

文章编号:1005-3085(2010)06-0959-08

基于序列和局部信息熵的蛋白质折叠速率预测模型*

高建召, 胡　刚, 王　奎, 沈世镒

(南开大学数学科学学院与LPMC，天津　300071)

摘　要: 正确预测蛋白质折叠速率对理解蛋白质的折叠机制非常重要。本文从AAindex数据库中的531种残基物理化学性质、序列长度信息和局部结构信息熵中筛选特征，从而提出了一个基于蛋白质序列信息的线性回归模型。针对三种折叠机制two-state，multi-state和mixed-state，用Jackknife验证模型，预测的折叠速率和实验验证的折叠速率相关系数分别为0.790，0.829和0.778。本文结果表明四阶局部结构信息熵和折叠速率有很高的负相关性；蛋白质的长度和蛋白质的折叠速率成反比关系；螺旋的含量会加快蛋白质的折叠过程。对two-state蛋白质β折叠的含量会减慢蛋白质的折叠过程；和其他模型相比，我们提出的线性回归模型具有输入参数少，计算简单，平均绝对误差小的优点。

关键词: 蛋白质折叠速率；基因序列的预测方法；局部结构信息熵；线性回归

分类号: **AMS**(2000) 68Q30　　　**中图分类号**: O236　　　**文献标识码**: A

1　引言

蛋白质折叠问题是计算生物学和生物信息学中的核心问题之一[1]。预测蛋白质折叠速率对理解蛋白质的折叠机制和分析蛋白质折叠的决定因素非常重要。蛋白质的折叠速率是用来描述蛋白质从变性状态恢复到天然结构的快慢。从实验角度观察蛋白质的折叠分为两种机制：一种是二态折叠(two-state)，是指蛋白质从变性状态到天然结构的过程中不需要经过中间状态。另外一种是多态折叠(multi-state)，是指蛋白质从变性状态到天然结构的过程中至少经过一种以上的中间状态[2]。通常来讲，二态折叠用来描述小蛋白的折叠机制，多态折叠用来描述大体积蛋白的折叠机制。传统的实验方法来研究蛋白质折叠的方法有光谱，质谱，核磁共振等方法。随着实验数据的积累，为我们用数学模型来预测蛋白质折叠速率创造了条件。

按照特征来分，预测蛋白质折叠模型可以分为三类：第一类，三级结构模型。通过描述蛋白质的拓扑结构如，相对接触距(relative contact order, CO)[3]、长接触距(long-range order, LRO)[4]、总接触距(total contact order)[5]和绝对接触距(absolute contact order)[6]等利用三级结构的方法。第二类，二级结构模型。通过真实和预测的二级结构来预测折叠速率，代表方法有二级结构含量的方法(SSC)[7]，通过从已知的二级结构或者预测的二级结构来预测折叠速率[8]。Ivankov和Finkelstein的有效折叠链长度模型，通过预测的二级结构计算有效长度(effective chain length, Leff)[9]来预测蛋白质折叠速率，近期的工作还有[10]等等。第三类，利用序列信息来构建模型。如Gromiha等利用氨基酸的49种物理化学属性来预测折叠速率，相关系数达到0.93等等。通过分析这些方法，我们发现这些方法的一个共同特点就是都利用了

收稿日期: 2009-01-09.　**作者简介**: 高建召(1983年3月生)，男，博士. 研究方向：生物信息学.

***基金项目**: 国家自然科学基金(10671100; 20836005)；刘徽应用数学研究中心、天津大学、南开大学联合研究项目；天津市自然科学基金(07JCZDJC06400).

蛋白质的二级或者三级结构信息建立模型。虽然Gromiha等的方法在计算中没有用到具体的结构信息，但是是建立在已知蛋白质结构类型的基础上进行预测的[1]。

本文考虑从氨基酸的物理化学性质出发，结合序列长度和局部结构信息熵[11]，来预测蛋白质的折叠速率。考虑到二态蛋白和多态蛋白的折叠机制有很大的差异，我们提出三个线性模型分别预测二态(two-state)，多态(multi-state)以及混合态(mixed-state，当我们不知道蛋白质的折叠机制时，蛋白质可能属于二态折叠或者多态折叠，我们称这种状态是混合态)的折叠速率。我们的方法主要分为以下几个步骤：首先提取蛋白质的序列特征，其次筛选出和折叠速率相关性比较高的特征，最后利用线性回归模型建模。

2 数据和方法

本文用到的数据集是已被实验证实的62个蛋白质的折叠速率数据集，该数据集曾被Ivankov和Finklstein使用，记为D62。该数据集包括，37个二态蛋白和25个多态蛋白。其中37个二态蛋白的平均长度为83，25个多态蛋白的平均长度为143，62个蛋白混合在一起的平均长度为107。这里的实验折叠速率是指实验观察数据的以十为底的对数$\log_{10}(f_k)$。该数据集可从网址http://mathbio.nankai.edu.cn/jzgao/folding_rate_database.htm下载。

2.1 实验设计

本文使用线性回归模型来拟合实验数据。线性回归方程表示为

$$\hat{y}=\sum_{i=1}^{n}\omega_i x_i+C,$$

其中$\hat{y}$为模型预测的折叠速率，x_i表示我们挑选的第i个特征，$i=1,\cdots,n$。n表示模型使用的特征个数，ω_i表示第i个特征x_i的回归系数，C是线性模型中的常数项。参数ω_i, C可以用最小二乘法求出。

我们利用Resubstition和Jackknife两种方法来检验模型。Resubstition是在训练集的基础上建立模型，并且用模型预测的训练集的折叠速率。然后计算预测的折叠速率和真实的折叠速率的相关系数。这个检验方法是来验证所建立的模型是否准确地描述了训练集。虽然这个方法很容易造成过拟合，但是先前描述的方法中[3,6,9,12]都曾用到过。为了便于和先前的方法做比较，我们也采用了这个检验方法。Jackknife方法，也称留一法。设训练集有n条观察数据，Jackknife检验用其中$n-1$条用来建立模型，剩下的1条数据用来检验模型，这样重复n次后，得到n个预测数值，然后和真实数据计算相关系数。本文即用Pearson相关系数(PCC, Pearson correlation coefficient)，平均绝对值误差(MAE, mean absolute error)作为评价指标，其定义如下

$$\mathrm{PCC}=\frac{\sum\limits_{i=1}^{N}(\hat{y}_i-\bar{\hat{y}})(y_i-\bar{y})}{\sqrt{\sum\limits_{i=1}^{N}(\hat{y}_i-\bar{\hat{y}})^2\sum\limits_{i=1}^{N}(y_i-\bar{y})^2}},\quad \mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i-y_i|,$$

其中$\hat{y}_i$是第i个蛋白质的预测折叠速率，y_i是第i个蛋白质的真实折叠速率，N表示该数据集中含有蛋白质的数目。$\bar{\hat{y}}$是所有预测折叠速率的平均值，其定义为

$$\bar{\hat{y}}=\frac{1}{N}\sum\hat{y}_i,$$

$\bar{y}$ 是所有真实折叠速率的平均值，其定义为

$$\bar{y} = \frac{1}{N}\sum y_i.$$

2.2 特征设计

AAindex 数据库[13] 记录了每种氨基酸各种不同的数值化属性，包括氨基酸的疏水性、体积和极性等等。我们从 AAindex 数据库中下载所有的氨基酸物理化学性质共 544 种。去除掉含有NA的数据 (残缺数据)，还剩下 531 种数据。这每一条数据含有 20 个数值，分别对应 20 种氨基酸的某种物理化学性质。这些数据按照公式

$$P_i^{\text{norm}} = \frac{P_i - P_{\min}}{P_{\max} - P_{\min}}$$

标准化到 0 到 1 之间。其中 P_i, P_i^{norm} 分别指第 i 种氨基酸的某种物理化学性质的数值和标准化后的数值，$P_{\max}$，$P_{\min}$ 分别是指这种物理化学性质的最大值和最小值。每条蛋白质的物理化学性质用平均含量来表示

$$P_{\text{ave}}^k = \sum_{i=1}^{L} P_i^k / L,$$

其中 L 为蛋白质长度，P_i^k 是蛋白质中第 i 个残基对应的第 k 个物理化学性质，k 是第 k 条 AAindex 记录信息，$k = 1, \cdots, 513$, $i = 1, \cdots, L$。

蛋白质的结构信息熵 (structural entropy) 是近些年 Chan 等[11] 提出的。它和蛋白质的热稳定性有很强的线性正相关关系。已经应用在蛋白质热稳定设计等方面[14,15]。我们分别计算每条蛋白质的三阶和四阶的平均局部结构信息熵 (local structural entropy, LSE)。设蛋白质长度为 n，共有 $n-3$ 个长度为 4 的片段，计算这 $n-3$ 个局部信息熵 (LSE) 的平均值，得到四阶的平均信息熵。三阶信息熵的定义类似。三阶，四阶残基片段对应的数值从这个链接 (http://sdse.life.nctu.edu.tw/index.cgi?xln=download) 下载，选择 scop-35-3-ss.txt 和 scop-35-4-ss.txt 这两个文件。更多的局部信息熵信息可以参考文献 [11]。我们还用到蛋白质序列长度 L 和长度的自然对数 $\ln(L)$ 作为特征。这样我们就为每条蛋白质建立一个 $531 + 2 + 2 = 535$ 维的特征向量。

2.3 特征选择

我们按照以下三个步骤来选择特征：

1) 利用 correlation-based feature subset selection (CFSS)[16] 方法剔除掉比较弱的特征；
2) 分别用向前和向后的方法来选择从步骤 1) 中得到的特征；
3) 从步骤 2) 中选择相关系数中最大的特征组合。

CFSS 方法近年来被成功应用在逻辑回归中[17]。CFSS 方法通过评价每个特征的预测能力给出优化的特征子集。我们将建立好的 535 个特征，利用 CFSS 方法做 10 折叠-交叉验证，当至少有一个 fold 的选择到这个特征时就保存该特征。这样第一步筛选后对应 two-state，mulit-state 和 mixed-state 的特征个数分别为 37, 432 和 60。第二步，分别按照向前的方法和向后的方法来选择特征。向前方法，就是向一个特征集合中添加一个特征，如果该特征能够提高模型的相关系数，就允许添加该特征。模型利用 Jackknife 检验方法做检验。计算每一个特征的线性回归模型，选择相关系数的最高的一个特征作为初始特征集合。然后添加新的一个特征如果相关系数增加，就保存这个特征。向后的方法，正好相反，利用所有的特征计算线性回归模型。如果去除掉某一个特征不会减少模型的相关系数，就去除掉该特征。模型利用 Jackknife 检验方法检验。用这两种方法搜索从第一步留下的特征，

最终对应two-state，mulit-state和mixed-state模型的特征个数分别为6, 5和6。所选的特征及描述如表1所示。我们发现选取的特征中都用到了长度和四阶的局部结构信息熵，没有用到三阶的局部结构信息熵。来自AAindex的特征都是关于蛋白质结构相关的特征。例如Norm_Freq_Beta是α/β类蛋白质中β折叠的正规化后的频数，Weight-Coil是用移动窗口下螺旋的权重等等。更具体的特征描述，可以参考AAindex数据库中对应ID下的描述。

表1: 三类线性回归模型选择的特征及相关系数

折叠机制	特征名称	AAindexID/特征描述	PCC
Two-state	Norm_Freq_Beta	PALJ810109	-0.639
	Ln_L	取自然对数后的长度	-0.483
	LSE4	四阶局部结构信息熵	-0.540
	AA_EXT	NAKH920103	-0.467
	VDW_Epsilon	LEVM760107	0.154
	AL	RACS820103	-0.430
Multi-state	L	蛋白质序列长度	-0.803
	Aver_Energy	OOBM850104	-0.098
	Part_Vol	BULH740101	0.107
	Weight_Coil	QIAN880131	0.214
	Relative_Mutable	DAYM780101	0.026
Mixed-state	Ln_L	取自然对数后的长度	-0.677
	Norm_Freq_Beta	PALJ810109	-0.497
	LSE4	四阶局部结构信息熵	-0.217
	AA_Mt_protein	NAKH900105	-0.167
	Relative_Mutable	DAYM780101	-0.012
	Aver_AL	RACS820103	-0.156

3 结果和讨论

本文预测蛋白折叠速率的模型基于三个线性回归模型。如果用户知道查询蛋白的折叠机制，就可以分别用二态(two-state)和多态(multi-state)的模型。如果用户不确定查询蛋白的折叠机制，可以利用mixed-state的线性模型。表2是我们得到的在数据集上分别预测two-state，multi-state和mixed-state态蛋白折叠速率的线性回归模型。从表2中我们可以看出各影响因子之间与折叠速率的相关关系。我们看到三个回归模型中长度跟折叠速率都成负相关关系。这和文献[9]中的结果吻合的很好。蛋白质越大，长度越长，需要折叠的时间就越多，折叠速率越慢。在two-state和mixed-state模型中，我们注意到四阶结构信息熵比三阶结构信息熵更有效并且四阶结构信息熵和折叠速率成负相关系数。我们知道结构信息熵和蛋白质的热稳定性有很好的正相关关系[11]。蛋白质越稳定，需要变性的温度T_m越高，从变性状态到天然结构状态折叠的速度越慢。在multi-state模型中，螺旋的含量(Weight_Coil)和

折叠速率成正相关，说明螺旋的含量的增多会加快蛋白质的折叠过程。对于two-state，β折叠(Norm_Freq_Beta)和折叠速率成负相关关系。这说明beta-strand阻碍了蛋白质的折叠过程，这个与文献[10]的结论是一致的。

表2: 蛋白质折叠速率预测模型

预测模型
Two-state
fold-rate_two=-9.301*Norm_Freq_Beta-1.629*Ln_L-14.819*LSE4-25.016*AA_EXT
-20.773*VDW_Epsilon-4.895*AL+56.684
Multi-state
fold-rate_multi=-0.01526*L+16.63188*Aver_Energy+28.29164*Part_Vol
+18.96604*Weight_Coil+11.91310*Relative_Mutable-33.76155
Mixed-state
fold-rate_mixed=-2.303*Ln_L-8.952*Norm_Freq_Beta-11.695LSE4-20.561AA_Mt_Protein
+18.671*Relative_Mutable-7.935Aver_AL+23.444

如果用Resubstitution检验方法检验模型，我们分别利用two-state，multi-state和mixed-state模型预测的折叠速率和真实折叠速率的PCC分别为0.855, 0.875和0.828。当用Jackknife检验方法检验模型，在two-state模型中，真实折叠速率和预测折叠速率的PCC为0.790，在multi-state和mixed-state模型中，PCC分别为0.829, 0.778。图1是我们用Jackknife检验方法检验模型，对三种折叠机制two-state，multi-state和mixed-state的预测折叠速率和真实的折叠速率的线性回归图。其中，三种折叠机制的预测结果和真实的折叠速率的PCC分别为0.790, 0.829和0.778。

为了检验我们的模型预测的效果，我们的模型和其他的十种模型进行了比较。这十种模型包括CO[3], LRO[4], TCD[5], ABS-CO[6], SSC[7], Leff[9], PPFR[10], CI[12], K-Fold[18]和QRSM[19]。表3和表4分别列出了用Resubstitution和Jackknife检验方法的结果。用Resubstitution的检验方法比较，我们的模型预测two-state，mulit-state和mixed-state的折叠速率的相关系数分别达到0.855, 0.875和0.828。在用Jackknife检验时，我们预测的三种状态的蛋白质折叠速率和实验验证的折叠速率的的相关系数分别为0.790, 0.829和0.778。从表4中我们可以看到，本文的模型的相关系数要低于PPFR模型。但是我们的Resubstitution检验和Jackknife检验方法预测结果的平均绝对误差(MAE)分别为0.595和0.729。这要比PPFR[10]方法在相同数据集上的Resubstitution检验和Jackknife检验的平均绝对误差0.88和0.93要低。另外，PPFR的模型输入参数比本文模型较多。例如，PPFR用于预测三种折叠态蛋白速率模型的参数分别为10, 10和8，而本文只用到6, 5和6。PPFR模型还用到PROTEUS[20]和PSIPRED[21]两个软件提供的预测的二级结构信息。我们的模型只利用了序列的信息和结构熵不需要额外的软件支持，模型参数更少，计算更简单。我们还注意到QRSM方法用Jackknife检验的结果也优于我们的结果。分析主要原因有，QRSM模型用二次响应曲面来回归49个特征，它的输入特征比我们的模型多，另外它的二次响应曲面模型比线性回归模型更加复杂。但QRSM方法只给出了预测mixed-state的模型，我们的方法优势在于给出不同折叠机制的预测蛋白质折叠速率的模型，输入特征都基于序列特征，而且输入特征个数较少，模型计算简单快捷。

表3和表4中部分数据来自文献[10,12]。表4中的K-fold方法数据来自文献[18]，该方法只给了mixed-state模型，用5倍交叉验证检验模型。

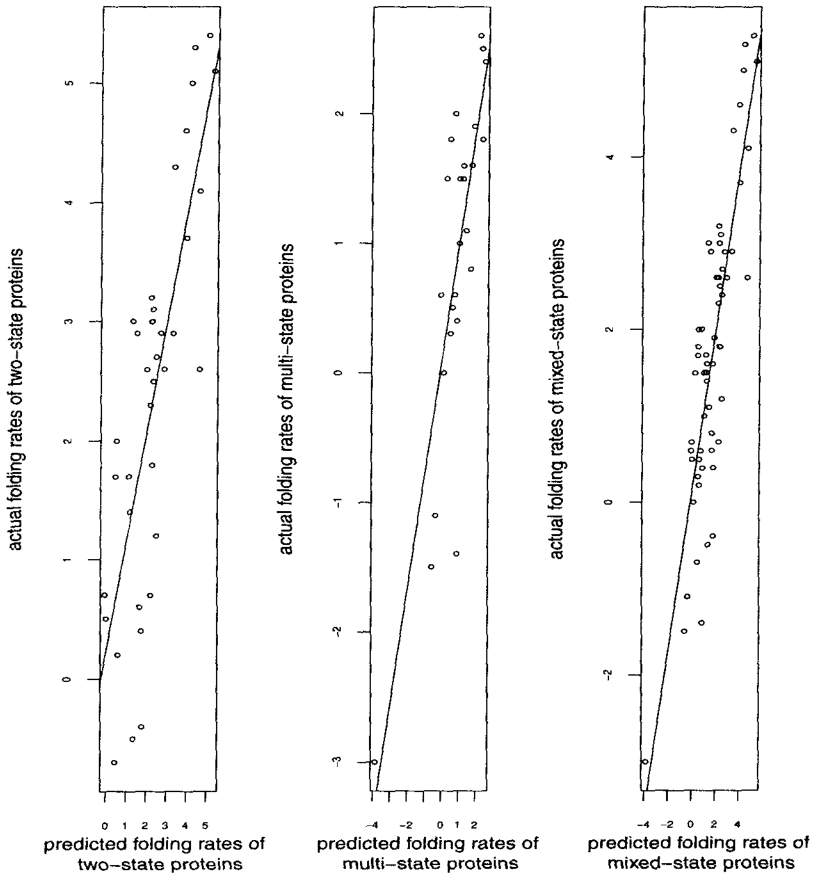

图1: 本文模型预测的折叠速率和真实折叠速率的线性回归图形(用Jackknife检验方法)

表3: 其他预测模型的比较(Resubstitution检验方法)

折叠机制	CO	ABS-CO	LRO	TCD	SSC	Leff	CI	PPFR	本文方法
Two-state	-0.57	-0.64	-0.79	-0.79	0.64	-0.61	0.73	0.92	0.855
Multi-state	0.435	-0.44	-0.34	0.23	-0.01	-0.88	0.70	0.92	0.875
Mixed-state	0.12	-0.57	-0.61	-0.19	0.42	-0.73	0.72	0.85	0.828

表 4: 其他预测模型的比较 (Jackknife 检验方法)

折叠机制	CI	K-Fold	QRSM	PPFR	本文方法
Two-state	0.73	N/A	N/A	0.87	0.790
Multi-state	0.70	N/A	N/A	0.87	0.829
Mixed-state	0.73	0.74	0.89	0.82	0.778

4 结论

我们的方法针对三种不同的蛋白质折叠机制，从 AAindex 数据库，序列长度信息和结构信息熵中筛选特征，给出了三个预测模型。我们结果表明：

1) 局部结构信息熵和蛋白质折叠速率成负相关，并且四阶局部信息熵要比三阶局部信息熵更有效；

2) 在 two-state 模型中，β 折叠的含量可能减缓蛋白质折叠过程；

3) 螺旋 (coil) 的含量能加速蛋白质的折叠过程。

我们模型预测的折叠速率和实验验证的折叠速率都达到 0.7 以上。和其他模型相比，我们的模型有需要的参数较少，计算简单，平均绝对误差小的优点。

参考文献：

[1] 郭建秀，马彬广，张红雨. 蛋白质折叠速率预测研究进展[J]. 生物物理学报，2006, 4(2): 89-95
Guo J X, Ma B G, Zhang H Y. Progress in protein folding rate prediction[J]. Acta Biophysica Sinica, 2006, 4(2): 89-95

[2] Jackson S E. How do small single-domain proteins fold[J]. Folding Design, 1998, 3(4): 81-91

[3] Plaxco K W, Simons K T, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins[J]. Journal of Molecular Bioolgy, 1998, 277(4): 985-994

[4] Gromiha M M, Selvaraj S. Comparison between long-range interactions and contact order in determining the folding rate of twostate proteins: application of long-range order to folding rate prediction[J]. Journal of Molecular Biology, 2001, 310(1): 27-32

[5] Zhou H Y, Zhou Y Q. Folding rate prediction using total contact distance[J]. Biophysical Journal, 2002, 82(1): 458-463

[6] Ivankov D N, *et al.* Contact order revisited: influence of protein size on the folding rate[J]. Protein Science, 2003, 12(9): 2057-2062

[7] Gong H, Isom D G, Srinivasan R, *et al.* Local secondary structure content predicts folding rates for simple, two-state proteins[J]. Journal of Molecular Biology, 2003, 327(5): 1149-1154

[8] Mimy L, Shalhnovich E. Protein folding theory: from lattice to all-atom models[J]. Annual Review of Biophysics and Biomolecular Structure, 2001, 30: 361-396

[9] Ivankov D N, Finkelstein A V. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure[J]. Proceedings of the National Academy of Sciences, 2004, 101(24): 8942-8944

[10] Jiang Y, Iglinski P, Kurgan L. Prediction of protein folding rates from primary sequence using hybrid sequence representation[J]. Journal of Computational Chemistry, 2009, 30(5): 772-783

[11] Chan C H, *et al.* Relationship between local structural entropy and protein thermostability[J]. Proteins: Structure, Function, and Bioinformatics, 2004, 57(4): 684-691

[12] Ma B G, Guo J X, Zhang H Y. Direct correlation between porteins' folding rates and their amino acid compositions: an Ab inito folding rate prediction[J]. Proteins: Structure, Function and Bioinformatics, 2006, 65(2): 362-372

[13] Kawashima S, *et al.* AAindex: amino acid index database, progress report 2008[J]. Nucleic Acids Research, 2008, 36: 202-205

[14] Bae E, *et al.* Bioinformatic method for protein thermal stabilization by structural entropy optimization[J]. Proceedings of the National Academy of Sciences, 2008, 105(28): 9594-9597

[15] Bannen R M, *et al.* Optimal design of thermally stable proteins[J]. Bioinformatics, 2008, 24(20): 2339-2343

[16] Hall M A. Correlation-based feature selection for machine learning[OL]. http://www.cs.waikato.ac.nz/mhall/thesis.pdf, 1999

[17] Landwehr N, Hall M, Frank E. Logistic model trees[J]. Machine Learning, 2005, 59(1/2): 161-205

[18] Capriotti E, Casadio R. K-Fold: a tool for the prediction of the protein folding kinetic order and rate[J]. Bioinformatics, 2007, 23(3): 385-386

[19] Huang L T, Gromiha M M. Analysis and predicton of protien folding rates using quadratic response surface models[J]. Journal of Computational Chemistry, 2008, 29(10): 1675-1683

[20] Montgomerie S, *et al.* Improving the accuracy of protein secondary structure prediction using structural alignment[J]. BMC Bioinformatics, 2006, 7: 301

[21] Jones D T. Protein secondary structure prediction based on position-specific scoring matrices[J]. Journal of Molecular Biology, 1999, 292(2): 195-202

Prediction Model of the Protein Folding Rate Using Sequence Representation and Local Structural Entropy

GAO Jian-zhao, HU Gang, WANG Kui, SHEN Shi-yi

(School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071)

Abstract: Prediction of protein folding rates is important in understanding the overall folding the mechanism. This article selects the features from 531 physical chemistry properties in the AAindex database, the length of proteins and the local structural entropy, and proposes three sequence-based linear regression models for two-state, multi-state and mixed-state proteins. The correlation between predicted folding rates and experimental folding rates for different folding kinetics is 0.790, 0.829 and 0.778, respectively. We show that the tetra-local structural entropy is negatively correlated with the protein folding rate. Length of protein is negatively correlated with the folding rates. Coil content may accelerate the protein folding process and for two-state proteins, the beta-strand content may decelerate the folding process. Compared with other models, our proposed method has advantages in less features, simple computation and smaller mean absolute errors.

Keywords: protein folding rates; sequence-based prediction; local structural entropy; linear regression

Received: 09 Jan 2009. **Accepted:** 22 Oct 2009.

Foundation item: The National Natural Science Foundation of China (10671100; 20836005); the Liuhui Center for Applied Mathematics, Joint Program of Tianjin and Nankai Universities; the Natural Science Foundation of Tianjin (07JCZDJC06400).